

Wasserstein GAN

Сергей Миллер

December 25, 2018

Contents

1 Введение	1
2 Метрика The Earth-Mover(EM) distance или Wasserstein-1	1
3 Оптимизация EM-distance	2
4 Алгоритм Wasserstein GAN	2
5 Генерация изображений с помощью WGAN	3
6 WGAN vs WGAN-GP	4

1 Введение

В данном докладе будет рассмотрена задача генерации изображений из распределения заданного выборкой(то есть набором изображений-примеров). Оказывается, что достаточно распространенной ситуацией является расположение изображений из целевого распределения на некотором низкоразмерном многообразии в общем пространстве признаков(пикселей изображения). В таких случаях методы, оптимизирующие правдоподобие параметризованного распределения:

$$\max_{\theta} \mathbf{E}_X \log P_{\theta}(x)$$

не имеют даже теоретических гарантий сходимости, и на практике, для преодаления этих ограничений приходится использовать зашумление компонент изображения(или иные техники). Новый подход к генеративным моделям предлагает специальную функцию потерь, которая не требует зашумления или иной предобработки, обладает меньшей чувствительностью к изменениям гиперпараметров, а также подвержена в меньшей степени взрывам и испарению градиентов(vanishing gradients). Большая часть информации взята из статей (1) и (2).

2 Метрика The Earth-Mover(EM) distance или Wasserstein-1

Ключевой идеей WGAN является использование специальной метрики (в пространстве распределений):

$$W(P_r, P_g) = \inf_{\gamma \sim \Pi(P_r, P_g)} \mathbf{E}_{(x,y) \sim \gamma} |x - y|$$

где $\Pi(P_r, P_g)$ - набор всех возможных распределений, маргинальными распределениями которых являются P_r и P_g . Метрика имеет смысл оптимальной стоимости переноса массы из x в y , для того, чтобы превратить распределение P_r в P_g .

В оригинальной статье (1) доказывается, что сходимость по EM-distance равносильна слабой сходимости(по распределению) в отличие от дивергенции Кульбака-Лейбнера, которая к тому же достаточно часто может быть неопределенна в окрестности искомого распределения.

3 Оптимизация EM-distance

Так как оптимизация EM-distance является достаточно трудной задачей, на практике решается в следующем виде: распределение P_θ задается параметризованной функцией g_θ от случайной величины $z \sim P_z$ (с известным распределением).

После такой параметризации использование следующих тождеств позволяет получить простой алгоритм оптимизации EM-distance:

$$W(P_r, P_\theta) = \sup_{\|f\|_L \leq 1} (\mathbf{E}_{x \sim P_r} f(x) - \mathbf{E}_{y \sim P_\theta} f(y))$$

1

$$\nabla W(P_r, P_\theta) = -\mathbf{E}_{z \sim p(z)} \nabla_\theta f^*(g_\theta(z))$$

2

где $\|\cdot\|_L$ - Липшицева норма, P_r - распределение реальных данных, а $f^* = \arg \max_f$ из тождества 1.

На практике это означает, что мы можем, используя например нейронные сети, задать g_θ и f_w , поочередно выполняя шаги градиентного спуска(соответствующий равенствам 1, 2), пока g_θ не приблизится достаточно P_θ к P_r , а f_w к f^* .

4 Алгоритм Wasserstein GAN

В итоге алгоритм в каноничном виде выглядит следующим образом:

Require: : α , the learning rate. c , the clipping parameter. m , the batch size.
 n_{critic} , the number of iterations of the critic per generator iteration.

Require: : w_0 , initial critic parameters. θ_0 , initial generator's parameters.

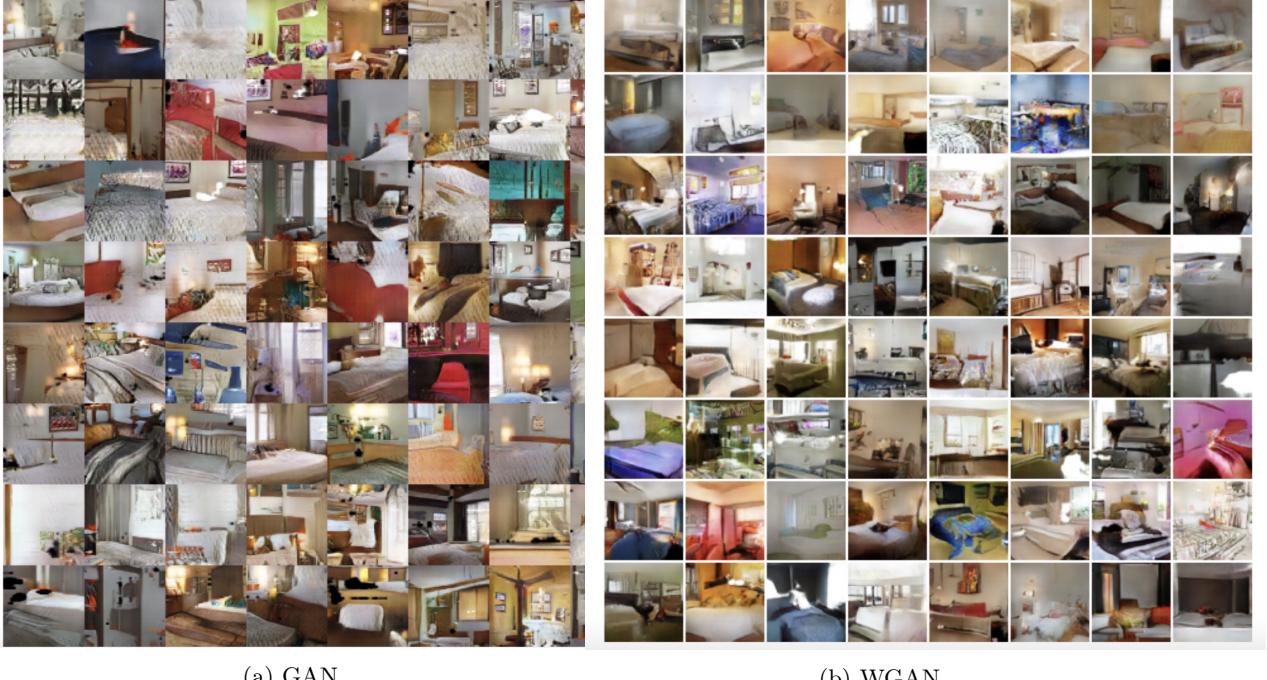
```

1: while  $\theta$  has not converged do
2:   for  $t = 0, \dots, n_{\text{critic}}$  do
3:     Sample  $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$  a batch from the real data.
4:     Sample  $\{z^{(i)}\}_{i=1}^m \sim p(z)$  a batch of prior samples.
5:      $g_w \leftarrow \nabla_w \left[ \frac{1}{m} \sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)})) \right]$ 
6:      $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$ 
7:      $w \leftarrow \text{clip}(w, -c, c)$ 
8:   end for
9:   Sample  $\{z^{(i)}\}_{i=1}^m \sim p(z)$  a batch of prior samples.
10:   $g_\theta \leftarrow -\nabla_\theta \frac{1}{m} \sum_{i=1}^m f_w(g_\theta(z^{(i)}))$ 
11:   $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$ 
12: end while
```

Стоит заметить, что важной частью реализации является искусственное ограничение всех весов сети дискриминатора f_w в диапазоне $[-c, c]$, для соблюдения условия Липшицевости оптимизируемой функции, чего хотелось бы избежать (более слабые ограничения в процессе обучения рассматривают в (2)).

5 Генерация изображений с помощью WGAN

Рассмотрим сгенерированные примеры изображений для WGAN и классической схемой генеративной сети в задаче генерации интерьера комнаты (в обоих вариантах используются общие генератор и дискриминатор):



(a) GAN

(b) WGAN

Несмотря на то, что по качеству изображения почти не отличимы, видно что классический GAN генерирует чуть-менее реалистичные интерьеры (в том плане, что предметы расположены относительно друг друга непривычным образом).

Слабым местом алгоритма обучения является искусственное загрубление весов сети дискриминатора для форсирования липшицевости. Более того, в (2) показано, что веса критика(дискриминатора) очень быстро насыщаются, и функция достигает предельной константы, не сумев хорошо приблизить искомое распределение.

WGAN-GP лишен этого недостатка, для этого используется модифицированная функция потерь:

$$L(\theta, w) = \mathbf{E}_{y \sim P_\theta} f_w(y) - \mathbf{E}_{x \sim P_r} f_w(x) + \lambda \mathbf{E}_{x \sim \frac{P_r + P_\theta}{2}} (\|\nabla_x f_w(x)\|_2 - 1)^2$$

3

Смысл такой функции потерь в начислении дополнительного штрафа за отклонение нормы градиентов функции f от 1, что в каком-то смысле ограничивает скорость роста функции f В (2) доказывается, что оптимизация такой функции потерь критика для WGAN также гарантирует липшицевость f_w но нет потери выученной информации в весах сети.

6 WGAN vs WGAN-GP

В задаче генерации интерьеров комнат качество сгенерированных изображений на порядок выше обычного WGAN, а также видно меньшее количество артефактов:

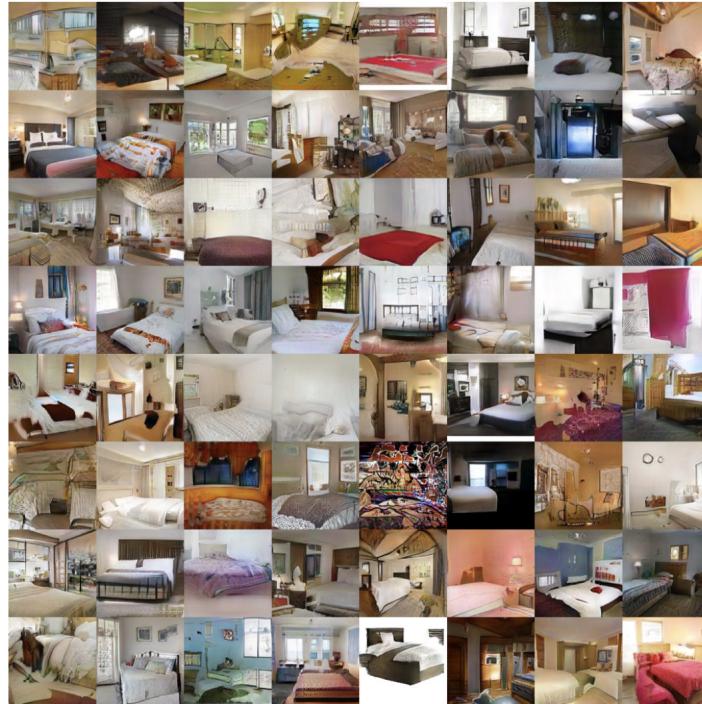


Figure 2: WGAN-GP (использована та же архитектура генератора и критика, что и в предыдущих примерах)

Также стоит отметить высокую стабильность при обучении WGAN-GP в различных конфигурациях архитектур генератора и критика по отношению к WGAN (а также DCGAN и LSGAN):

DCGAN	LSGAN	WGAN (clipping)	WGAN-GP (ours)
Baseline (G : DCGAN, D : DCGAN)			
G : No BN and a constant number of filters, D : DCGAN			
G : 4-layer 512-dim ReLU MLP, D : DCGAN			
No normalization in either G or D			
Gated multiplicative nonlinearities everywhere in G and D			
tanh nonlinearities everywhere in G and D			
101-layer ResNet G and D			

References

- [1] Martin Arjovsky, Soumith Chintala, Léon Bottou *Wasserstein GAN*
- [2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville *Improved Training of Wasserstein GANs*